Responses of human frontal cortex to surprising events are predicted by formal associative learning theory

P. C. Fletcher[1], J. M. Anderson[1], D. R. Shanks[2], R. Honey[3], T. A. Carpenter[4], T. Donovan[4], N. Papadakis[4,5] and E. T. Bullmore[1,4]

[1] *Brain Mapping Unit, Box 189, Department of Psychiatry, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK*

[2] *Department of Psychology, University College London, London WC1E 6BT, UK*

[3] *Center for Clinical Research in Neuropsychiatry, University of Western Australia, Perth, Western Australia, Australia*

[4] *Wolfson Brain Imaging Center, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK*

[5] *Sheffield University, Sheffield, S3 7RH, UK*

Correspondence should be addressed to P.F. (pcf22@cam.ac.uk)

Published online: 17 September 2001, DOI: 10.1038/nn733

Learning depends on surprise and is not engendered by predictable occurrences. In this functional magnetic resonance imaging (fMRI) study of causal associative learning, we show that dorsolateral prefrontal cortex (DLPFC) is associated specifically with the adjustment of inferential learning on the basis of unpredictability. At the outset, when all associations were unpredictable, DLPFC activation was maximal. This response attenuated with learning but, subsequently, activation here was evoked by surprise violations of the learned association. Furthermore, the magnitude of DLPFC response to a surprise event was sensitive to the relationship that had been learned and was predictive of subsequent behavioral change. In short, the physiological response properties of right DLPFC satisfied specific predictions made by associative learning theory.

Our exquisite sensitivity to associative relationships in the environment is demonstrated by an ability to evaluate their strength with great accuracy and to adapt this evaluation rapidly in response to unexpected occurrences[1]. This attribute has been extensively studied behaviorally in humans[2] but remains to be characterized at the neurophysiological level. Associative learning theory suggests that unpredictability—a discrepancy between a predicted and an actual outcome—forms the basis for learning[3]. This view, embodied in major current connectionist theories of learning[4,5] and validated at an electrophysiological level[6], accounts for a number of behavioral phenomena noted in associative learning experiments[1,7]. It may be formulated in the rule of Rescorla and Wagner[8]:

$$\Delta V_{\text{cue}} = \alpha\beta\,(\lambda - \Sigma V) \qquad (1)$$

Here, ΔV_{cue} denotes the change in associative strength of a cue, which is modulated by two parameters: α, the learning parameter associated with the cue, and β, the learning parameter associated with the outcome. The maximum associative strength attainable is denoted by λ. ΣV is the associative strength of all cues present on a particular trial. Put simply, the change that a given cue–outcome pairing produces in the subjective association is expressed as a function of an error signal $(\lambda - \Sigma V)$, which arises from the violation of an expectancy. The learning process may be considered as the detection of a mismatch between predictions and outcomes together with an adjustment of expectancy on the basis of this mismatch. This formulation generates predictions that are testable with functional neuroimaging.

In this fMRI study of causal associative learning, we used a trial-specific design[9,10] to characterize brain responses to unpredictable (maximal learning) events occurring relatively rarely against a background of frequent, predictable (minimal learning) events. Subjects learned various associations between cues (fictitious drugs) and outcomes (fictitious syndromes) (**Fig. 1**). With learning established, brain responses to surprise events were determined. Because different types of cue–outcome (drug–syndrome) relationships were learned ('drug' is followed by 'syndrome'; 'drug' is not followed by 'syndrome'; 'drug' has no predictable relationship with 'syndrome'), we were able to characterize the modulation of surprise-dependent brain regions as a function of the nature of the learned relationship.

The Rescorla–Wagner learning rule generates four specific hypotheses with regard to identifying a brain region that reflects learning. First, there should be a change in level of event-related activity across the initial learning period (greater activation in early trials when every event is unpredictable). Second, subsequent to learning, events that violate a learned expectancy should produce activation in a learning-related region. A third prediction concerns the effects of the type of causal relationship that has been learned. In the positive con-

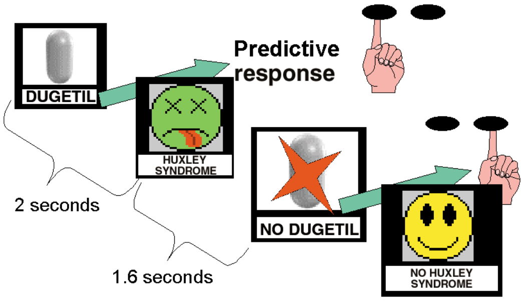

Fig. 1. Experimental procedure.

tingency, when the presence of 'drug' is a strong predictor of 'syndrome,' a surprise event is 'drug–no syndrome.' Under a learned negative contingency, 'drug–syndrome' is unexpected. According to the Rescorla–Wagner rule, these two types of unexpectedness should induce different adjustments in ΔV_{cue}. This follows because although the learning rate parameter α is determined by the cue, which is consistent under both contingencies, the learning rate parameter β is determined by the outcome, which differs between contingencies (because the outcome ('syndrome') is present in one case (negative contingency surprise event) but not in the other). In brief, the rule, supported by behavioral evidence[11], predicts a larger learning change for surprising cue–outcome pairings than for surprising cue–no outcome pairings. The fourth prediction is related to this. Within the setting of the same learned relationship, the learning engendered by unpredictable events will be modulated by the configuration of the event. Thus, when a negative association has been learned by seeing many 'drug–no syndrome' and 'no drug–syndrome' pairings, then the two rare surprise events will be 'drug–syndrome' and 'no drug–no syndrome' pairings. According to the Rescorla–Wagner rule, the former will have a greater influence upon learning because it comprises both cue and outcome. Thus, a direct comparison of these two conditions should also identify learning-related brain activation.

Our goal was to identify brain regions showing a highly specific pattern of activation (initially high, reducing with learning, re-evoked by unexpectedness, predictive of behavioral change). The subsequent series of analysis showed that right DLPFC met all of these criteria.

Results

Subjects' predictive responses show that they were sensitive to the causal relationship in each case (**Fig. 2**). In keeping with existing behavioral data[11], the positive causal relationship ('drug' is associated with 'syndrome') produced a significantly greater behavioral effect. In addition, as expected, when there was a violated expectancy in the setting of a learned negative relationship, subjects were more likely to show a behavioral change (to predict that 'drug' will be associated with 'syndrome') on the next trial than when expectancy was violated in the setting of a learned positive relationship ($p < 0.05$). Although subjects were sensitive to both positive and negative contingencies, the positive association was learned more strongly and was more robust in the face of unexpected trials.

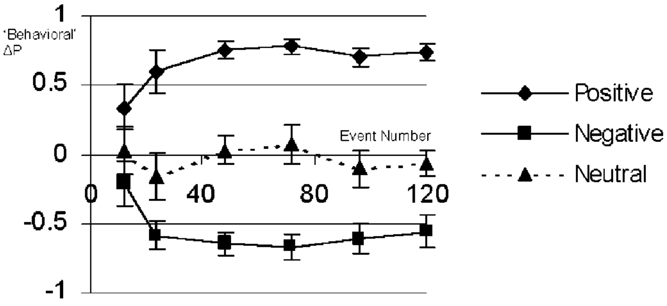

Fig. 2. Subjects' predictive tendencies on the three types of associative relationship. Behavioral ΔP values (with standard error bars) calculated as described in Methods section. Plots are averaged across the 11 subjects. ΔP values for individual subjects at each time point were calculated on the basis of responses across the 12 preceding items (during learning) or the 24 preceding items (once learning was established).

With regard to the neuroimaging data (**Fig. 3**), during the initial learning phase (across the first 15 trials) BOLD responses in bilateral frontal cortex attenuated quickly. Following learning, surprise trials of all types (in both negative and positive contingencies) compared to predictable trials produced activation of right DLPFC. The same region was also differentially sensitive to surprising outcomes to cues under the negative and positive contingencies. Evoked responses here were greater for the unpredicted 'drug–syndrome' events in the negative contingency than for the unpredicted 'drug–no syndrome' events in the positive contingency. Finally, the same region was also differentially sensitive to the two types of unpredictable events within the negative contingency condition: 'drug–syndrome' events produced greater right DLPFC activation than 'no drug–no syndrome' events, though both were equally rare and both occurred against the background of the same learned associative relationship.

Discussion

We identified a region of right DLPFC that uniquely satisfies our theoretically principled criteria for involvement in learning. The learning-related attenuation in activation, the recurrence of activation in response to surprise and the modulation of surprise-related activation by the nature of the surprising event show that physiological activation of this region may be accounted for by associative learning theory.

The finding that bilateral frontal regions were sensitive to initial learning is consistent with previous studies showing that activation in areas of frontal cortex decreases and disappears[12,13] as tasks become well-learned and automatic. Our second observation, that right DLPFC is sensitive to unpredictability, is also consistent with previous work[14,15]. We are confident for two reasons that higher error rates associated with unpredictable events do not explain right DLPFC activation. First, our experimental design included a neutral relationship condition by which we were able to control for the non-specific effects of error. Second, the subsequent analyses compared the effects of different types of surprise events. Error rates did not differ across these different events and cannot be invoked to account for the activation differences.

The most specific prediction of the Rescorla–Wagner rule, and the one that our study was expressly designed to explore at the functional neuroanatomical level, is that learning effects, as well as being engendered by surprise, are modulated by the configuration of the surprise event. This effect was seen in our behavioral data and in the right DLPFC response to surprise. Of course, it might be argued that the augmented response of

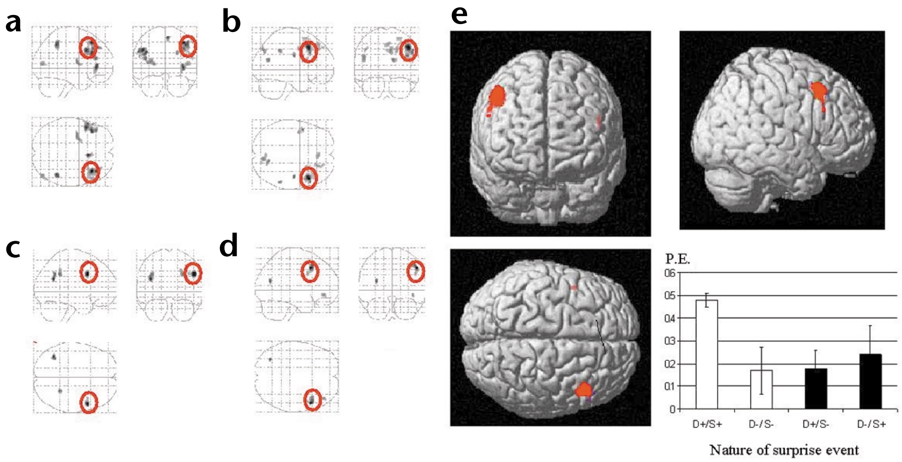

Fig. 3. Functional neuroimaging findings. The unmasked analyses are presented as maximum intensity projections or 'glass brain' figures, thresholded at p < 0.005, uncorrected for multiple comparisons. Details of the activations are given in Table 1. (**a**) 'Glass brain' figure for analysis testing prediction 1, decreases in activity during initial learning. (**b**) 'Glass brain' figure for analysis testing prediction 2, main effects of all unpredictable events. (**c**) 'Glass brain' figure for analysis testing prediction 3, modulation of unpredictability-related responses by type of causal relationship. (**d**) 'Glass brain' figure for analysis testing prediction 4, effects of different unpredictable events within the same learning session. (**e**) The masked analysis showing activation common to contrasts 1, 2, 3 and 4, superimposed upon a representation of a brain rendered into the same stereotaxic space. The accompanying graph shows the size of response (magnitude of parameter estimate (P.E.) from a voxel in right DLPFC for the different types of unpredictable events relative to the background predictable events) in the two learning conditions. Error bars show the standard error of the parameter estimate size across the subjects. White bars represent negative contingency; black bars, positive contingency. D, disease; S, syndrome. Thus, the greatest effect occurs in D+/S+ ('drug–syndrome') events with all other events producing significant though lesser effects.

right DLPFC to surprise in the negative relationship occurred because the overall level of associative strength (as indicated by subjects' predictive responses) was less in the negative than the positive contingency (**Fig. 2**). However, precisely the same region showed greater activation in response to 'drug–syndrome' than 'no drug–no syndrome' events when the analysis was confined to the negative relationship.

Our observation that DLPFC is sensitive to unpredictability—but not directly to error—is consistent with views that relate DLPFC and anterior cingulate cortex function to conditions in which conflict occurs[16,17]. This has been further demonstrated in a 'flanker' task[18,19], wherein a central cue, governing the nature of a response, is flanked by cues that may or may not be compatible with it. Anterior cingulate cortex and right DLPFC are sensitive to a response cue that is incompatible with its surroundings and to a configuration that violates an expectancy based on preceding trials[18]. It was suggested that error or conflict detection is associated with anterior cingulate and conflict resolution with lateral PFC activation. Our data are consistent with this model of the role of lateral PFC. Moreover, we are able to address a residual ambiguity of such studies, namely that reported activations may be sensitive to the relative rarity of the unexpected event. Thus, in this study[18], a 'conflict' event was one that had been preceded by a number of trials of a different type. Could it be that DLPFC is responsive merely to detecting this rarity rather than responding to the conflict? In our study, the inclusion of different types of unexpected events (each occurring with the same low frequency) enables us to rule this out.

Our data also support an electrophysiological study of cued motor responses in patients with lateral PFC damage who show a normal evoked electrical response to errors but weakened tendency to use this information to correct errors[20]. This suggests that lateral frontal cortex is associated not with error-monitoring/detection but with resultant compensatory mechanisms. Further evidence for this derives from the observation that, in the Stroop interference task, alerting a subject to imminent interference produces a preparatory activation in left DLPFC compared to the unprepared state[21]. Additionally, subjects with greatest DLPFC activation show the least effects of subsequent Stroop interference. This observation suggests a role for DLPFC in control implementation. Our results are consistent with this and suggest that the changes implemented in association with DLPFC activation extend beyond the current event to subsequent trials. That is, they can be considered as learning effects: an observation that is consistent with encoding studies showing PFC activation to be predictive of subsequent recall success[22,23].

Other interpretations of DLPFC activation have been proposed. In a 'go/no go' task in which both 'go' and 'no go' trials are equally frequent[24], bilateral DLPFC activity is greater in response to 'no go' trials than 'go' trials, suggesting that activity reflects response inhibition. However, lateral PFC was not activated in 'no go' trials relative to baseline, and a problem with this task is that 'go' and 'no go' trials may be differentiated according to whether or not a response occurred. In our study, this was not a problem because the response (common to all events) was a predictive rather than a reactive one. Thus, it is highly unlikely that unpredictable trials involved any response inhibition (because an unexpected event was designated as such by what occurred after the response had been made).

Non-automatic, novel tasks maximize learning[25,26] and the current results provide an example of how DLPFC function may contribute to this process. Specifically, activity seems to reflect the way in which subjects re-evaluate learned relationships or expectancies in response to unpredictability. It has been suggested that unpredictability is a *sine qua non* for learning[3], and a physiological correlate of this has been established in dopaminergic neurons of substantial nigra and ventral tegmentum in macaques[5]. Unless outcome violates expectancy, then learning parameters are not adjusted. We suggest that the pattern of activity observed here in right DLPFC makes this region a highly plausible candidate for a key locus of surprise-dependent learning in humans. This view resonates with the suggestion that a primary function of frontal cortex is in the provision of bias signals guiding activity in other brain regions in order to establish mappings between stimuli and appropriate responses[27]. Such signals would be maximized by unpredictable events and attenuated as predictability was established. Because they result in the establishment and consolidation of new pathways, they would predict a change in subsequent responses, a suggestion borne out by the results of the current study. A related though more anatomically specific position is that a key role of DLPFC lies in the mediation of cross-temporal contingencies[28,29]. DLPFC subserves processes required when the nature of a response is contingent upon a retained sensory cue. These formulations of prefrontal function in terms of an effect on response bias and cross-temporal contingencies are highly compatible with each other and with the results of the current study.

Table 1. fMRI activations.

Region	Coordinates[34]			Z score
(a) Prediction 1. Decreases in activity during initial learning.				
	x	*y*	*z*	
R SFG/MFG	**38**	**26**	**50**	**3.8**
	40	**28**	**38**	**3.1**
	46	**22**	**48**	**3.0**
	6	20	32	3.6
L MFG	–46	16	36	3.6
	–42	26	30	3.4
R parietal lobe	44	–48	46	3.4
R occipital lobe	46	–78	8	2.7
L putamen	–24	–2	8	2.9
(b) Prediction 2. Main effects of all unpredictable events.				
R MFG	**46**	**18**	**42**	**4.3**
	52	**22**	**34**	**3.6**
R SFG	22	48	44	2.8
L MFG/IFG	–48	6	26	3.5
	–50	14	44	3.1
R premotor cortex	44	–12	28	3.6
R inferior parietal lobe	56	–42	38	3.5
R precuneus	10	–66	20	3.5
(c) Prediction 3. Modulation of unpredictability-related responses by type of causal relationship.				
R MFG	**52**	**10**	**34**	**3.6**
	40	**10**	**36**	**2.6**
L temporal/	–38	–60	24	3.4
inferior parietal lobe	–36	–56	34	3.0
Cingulate gyrus/ white matter	26	–36	34	3.0
(d) Prediction 4. Effects of different unpredictable events within the same learning session.				
R MFG	**52**	**16**	**44**	**4.1**
	40	**10**	**34**	**3.0**
	42	42	–8	3.1
L temporal lobe	–32	–68	18	3.6

Coordinates of activation foci together with *Z* scores and an estimate of where the activations lie in anatomical terms are presented for each of the four contrasts. The activation foci highlighted in bold type (right DLPFC) are those which were common to all four contrasts. For completeness, other activations (including those at lower thresholds, $p < 0.01$) are reported. These are provided for information and we refrain from drawing any conclusions about them with regard to associative learning. SGF, superior frontal gyrus; MFG, middle frontal gyrus; R, right; L, left.

Methods

Twelve right-handed volunteers (6 female) with a mean age of 29 ± 8.8 years (± s.d.) were scanned. Subjects with a history of psychiatric or neurological illness, or of head injury or substance abuse were excluded. Participants had a mean predicted IQ of 110 ± 5.8 based upon the National Adult Reading Test[30]. The study was approved by the Addenbrooke's NHS Trust Local Research Ethics Committee and written informed consent was obtained from all subjects. One of the subjects (female) was subsequently found to be insensitive to the psychological manipulation and was excluded from further analysis, leaving data from 11 subjects.

Before entering the scanner, subjects were told that they should imagine that they were working for a drug company and would be required to determine the likelihood with which fictitious drugs would predict fictitious syndromes. For a given 'drug–syndrome' pair, they would be presented with 'case studies' and should use this information to learn whether the drug was likely to predict the syndrome. For each case study, they would be informed graphically whether or not the drug was administered and must then predict whether the syndrome would occur. After a couple of seconds, they would be told whether or not the syndrome occurred and the next 'case study' would then be presented. They must try to use successive case studies to make their predictions as accurate as possible. They were then given a short period of practice.

Each subject underwent 3 successive 7-min scanning sessions. When scanning began, subjects were presented visually with successive 'case studies' (**Fig. 1**), each requiring a prediction of whether the outcome (syndrome) would occur on the basis of whether the cue (drug) was present. Stimuli were presented using DMDX (K.I. Forster and J.C. Forster, Univ. of Arizona) on a screen placed comfortably within the subject's field of view. Each cue was presented for 2 s during which time the subject indicated their prediction with a button push. The cue then disappeared to be replaced by an icon representing the presence or absence of a syndrome and this stayed on the screen for 1.6 s. The next 'case' followed immediately. During each scanning session, 120 cases were presented, of which 25% were unexpected on the basis of the previous learning. Three fictitious drug names (Dugetil, Batatrim and Aubina) and three fictitious syndromes (Hamkaoman, Huxley and Lyndsay) were taken from a behavioral study[31]. For a single session, only information with respect to one drug and one syndrome was presented. Across the three sessions, therefore, a subject was exposed to three different pairings. The nature of these pairings (that is, which cues occurred with which syndromes and which pairings were used to illustrate a negative or positive causal relationship) was counterbalanced across subjects.

Unknown to subjects, each session carried a different causal relationship between cue and outcome. These relationships are expressed in terms of ΔP:

$$\Delta P = P(\text{'syndrome' following 'drug'}) - P(\text{'syndrome' following 'no drug'}) \qquad (2)$$

ΔP is therefore simply a measure of the extent to which the presence of the drug alters the probability of the occurrence of the syndrome.

The first relationship was a positive relationship: $\Delta P = P$('syndrome' following 'drug') – P('syndrome' following 'no drug') = 0.75 – 0.25 = +0.5. The second was a negative relationship: $\Delta P = 0.25 - 0.75 = -0.5$. The third was a neutral relationship in which $\Delta P = 0.5 - 0.5 = 0$. Order of corrections was varied across subjects.

Scanning. Imaging data were collected using a Bruker Medspec (Ettlingen, Germany) scanner operating at 3 Tesla. 151 T2*-weighted echo-planar images, depicting BOLD contrast, were acquired in each session (TE, 27 ms; TR, 3.1 s). Twenty-one slices (each of 4 mm thickness; interslice gap, 5 mm; matrix size, 128 × 128) per image were acquired. The first 6 EPI images in each session were subsequently discarded to avoid T1 equilibration effects, leaving 145 volumes per session. (In two subjects, data from one of the sessions were lost for technical reasons. In both cases, the missing data were acquired in the neutral learning condition. The remaining data for these subjects were used in the subsequent analysis of learning effects).

Behavioral analysis. A measure of subjects' evolving causal inference derives from their predictive responses to cues during scanning. An average 'behavioral' Δ*P* value for each session was derived from trials occurring after the first 15 events, a number chosen (on the basis of behavioral piloting) as the minimum at which learning would have been established. This behavioral Δ*P* value is based not upon the objective probabilities but rather upon the probabilities that subjects would predict the syndrome given the presence and the absence of the drug. The mean values for behavioral Δ*P* in the three conditions—positive relationship, negative relationship and neutral relationship—were 0.77 ± 0.3, –0.56 ± 0.4 and –0.05 ± 0.5, respectively. The magnitude of effects in the negative and positive relationship conditions differed significantly (paired *t*-test, *d.f.* = 10, $p < 0.05$). These data, averaged across sessions, provide only a broad idea of subjects' patterns of responses. To provide a clearer indication of subjects' behavior, each session was divided into 6 blocks (consisting of 12, 12, 24, 24, 24 and 24 events respectively) and Δ*P* for each block was calculated (**Fig. 2**).

We analyzed effects of unexpected events upon subjects' re-evaluation of the learned causal relationship. The effect of the unexpected event was assessed and compared across negative and positive associative relationships. As predicted, the magnitude of probability change generated by the average unpredictable trial in the negative relationship (mean ± s.e.m., +0.25 ± 0.08) exceeded (paired t test, DF, 10 $p < 0.05$) that occurring following unpredictability in the positive learning session (mean ± s.e.m., –0.09 ± 0.03).

Analysis of fMRI data. All data analysis was done using statistical parametric mapping[32] in the SPM99 program (Wellcome Department of Cognitive Neurology, London, UK). This included slice acquisition time correction, within-subject image realignment, spatial normalization to a standard template[33] (C.A. Cosoco *et al.*, *Neuroimage* **5**, S425, 1997) and spatial smoothing using a Gaussian kernel (8 mm full-width at half-maximum).

The time series in each session was high-pass filtered (to a maximum of 1/120 Hz). Events were designated as occurring at the presentation of the outcome stimulus. Four event types were modeled: 'drug–syndrome,' 'no drug–syndrome,' 'drug–no syndrome' and 'no drug–no syndrome.' In the case of the negative relationship, 'drug–syndrome,' and 'no drug–no syndrome' were the unexpected events. In the case of the positive relationship, 'drug–no syndrome' and 'no drug–syndrome' were unexpected. The other events in each case formed the expected events with which these were contrasted. In the neutral condition, each of these event-types was modeled along with a specification as to whether it had been correctly or incorrectly predicted. For each of the sessions, the first 15 events were designated as the learning phase and defined by their own partition in the design matrix.

The average hemodynamic responses to each event type were modeled using a canonical, synthetic hemodynamic response function[10]. This function was used as a covariate in a general linear model and a parameter estimate was generated for each voxel for each event type. The parameter estimate, derived from the mean least squares fit of the model to the data, reflects the strength of covariance between the data and the canonical response function for a given condition. Individuals' contrast images, derived from pair-wise contrasts between parameter estimates for different events, were taken to a second level group analysis in which *t*-values were calculated for each voxel treating inter-subject variability as a random effect. The *t*-values were transformed to unit normal *Z* distribution to create a statistical parametric map for each of the planned contrasts. Contrasts were thresholded at $p < 0.001$, uncorrected for multiple comparisons. An uncorrected threshold was chosen as the serial masking procedure makes whole brain corrections inappropriate.

Contrasts. Prediction 1 was that activity in learning-related areas would decrease during initial learning. Learning across the initial 15 trials was modeled as a linear decrease in magnitude of evoked response from the first trial (in which learning should be maximal) to the fifteenth, in which a high level of predictability should engender minimal learning (**Table 1a**, **Fig. 3**, regions showing such a linear attenuation). They are right and left dorsolateral PFC, occipital and parietal cortex and left putamen. Results of this analysis were used as a 'mask' within which we explored the effects of unpredictable events occurring during the remaining 105 trials.

Prediction 2 was that learning-related regions should show main effects of all unpredictable events. Across both positive and negative learning sessions, unpredictable events were contrasted with predictable events. Acknowledging that this comparison is partly confounded by correct versus incorrect predictions, we attempted to remove the nonspecific responses to error-related feedback by introducing data from the zero-learning contingency into the interaction. Thus, the contrast was set up as follows: ('unexpected' versus 'expected')$_{\text{NEG+POS}}$ versus ('incorrect' versus 'correct')$_{\text{NEUTRAL}}$. The only area of overlap with the first contrast above is right DLPFC (**Table 1b**, **Fig. 3**).

Prediction 3 was that in learning-related regions, unpredictability-related responses would be modulated by the type of causal relationship. We characterized regions in which the response to unpredictable ('drug–syndrome') events in the negative contingency was greater than the response to unpredictable ('drug–no syndrome') events in the positive contingency. The masked analysis showed overlap with the first and second comparisons in right DLPFC (**Table 1**, **Fig. 3**).

Prediction 4 was that, in learning-related regions, different unpredictable events have different effects within the same learning session. The fourth contrast explored the differences between these two types of events solely within the negative contingency learning session. In this session, both 'drug–syndrome' and 'no drug–no syndrome' events were surprising and occurred with equal rarity. Differences in BOLD responses evoked by these two types of event were found, within the masked region, in right DLPFC (**Table 1**, **Fig. 3**).

Acknowledgements

We thank H. Matute, F. Arcediano and R.R. Miller for allowing us to use their drug and syndrome names. We also thank J.D. Pickard and members of the Wolfson Brain Imaging Center, Addenbrooke's Hospital, Cambridge. P.C.F. is supported by the Wellcome Trust. D.R.S. is supported by the UK Biotech and Biological Sciences Research Council.

Received 17 July; Accepted 5 September 2001

1. Shanks, D. R. *The Psychology of Associative Learning* (Cambridge Univ. Press, Cambridge, UK, 1995).
2. Shanks, D. R., Medoff, D. R. & Medin, D. L. *The Psychology of Learning and Motivation: Causal Learning* (Academic, San Diego, 1996).
3. Schultz, W. & Dickinson, A. Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* **23**, 473–500 (2000).
4. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457 (1995).
5. O'Reilly, R. C. & Rudy, J. W. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychol. Rev.* **108**, 311–345 (2001).
6. Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43–48 (2001).
7. Dickinson, A. Causal learning: an associative analysis. *Quart. J. Exp. Psychol. B* **54**, 2–25 (2001).
8. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* (eds. Black, A. H. & Prokasy, W. F.) 64–99 (Appleton Century Crofts, New York, 1972).
9. Buckner, R. L. Event-related fMRI and the hemodynamic response *Hum. Brain. Mapp.* **6**, 373–377 (1998).
10. Friston, K. J. *et al.* Event-related fMRI: characterizing differential responses. *Neuroimage* **7**, 30–40 (1998).
11. Wasserman, E. A., Elek, S. M., Chatlosh, D. L. & Baker, A. G. Rating causal relations: the role of probability in judgments of response-outcome contingency. *J. Exp. Psychol. Learn. Mem. Cogn.* **19**, 174–188 (1993)
12. Raichle, M. E. *et al.* Practice-related changes in human brain functional anatomy during nonmotor learning. *Cereb. Cortex* **4**, 8–26 (1994).
13. Fletcher, P. C., Shallice, T. & Dolan, R. J. "Sculpting the response space"- an account of left prefrontal activation at encoding. *Neuroimage* **12**, 404–417 (2000).
14. Zeki, S. & Marini, L. Three cortical stages of color processing in the human brain. *Brain* **121**, 1669–1685 (1998).
15. Fink, G. R. *et al.* The neural consequences of conflict between intention and

the senses. *Brain* **122**, 497–512 (1999).

16. Menon, V., Adleman, N. E., White, C. D., Glover, G. H. & Reiss, A. L. Error-related brain activation during a Go/NoGo response inhibition task. *Hum. Brain Mapp.* **12**, 131–143 (2001).
17. Carter, C. S. *et al.* Anterior cingulate cortex, error detection and the online monitoring of performance. *Science* **280**, 747–749 (1998).
18. Casey, B. J. *et al.* Dissociation of response conflict, attentional selection and expectancy with functional magnetic resonance imaging. *Proc. Natl. Acad. Sci. USA* **97**, 8728–8733 (2000).
19. Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S. & Cohen, J. D. Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* **402**, 179–181 (1999).
20. Gehring, W. J. & Knight, R. T. Prefrontal cingulate interactions in action monitoring. *Nat. Neurosci.* **3**, 516–520 (2000).
21. MacDonald, A. W., Cohen, J. D., Stenger, V. A. & Carter, C. S. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* **288**, 1835–1838 (2000).
22. Wagner, A. D. *et al.* Building memories: remembering and forgetting of verbal memories as predicted by brain activity. *Science* **281**, 1188–1191 (1998).
23. Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H. & Gabrieli, J. D. Making memories: brain activity that predicts how well visual experience will be remembered. *Science* **281**, 1185–1187 (1998).
24. Liddle, P. F., Kiehl, K. A. & Smith, A. M. Event-related fMRI study of response inhibition. *Hum. Brain Mapp.* **12**, 100–109 (2001).
25. Sussman, G. J. *A Computational Model of Skill Acquisition* (American Elsevier, New York, 1975).
26. Shallice, T. *From Neuropsychology to Mental Structure* (Cambridge Univ. Press, Cambridge, UK, 1988).
27. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
28. Fuster, J. M. *The Prefrontal Cortex. Anatomy, Physiology and Neuropsychology of the Frontal Lobe* (Lippincott–Raven, Philadelphia, 1997).
29. Fuster, J. M. Executive frontal functions. *Exp. Brain Res.* **133**, 66–70 (2000).
30. Nelson, H. E. *The National Adult Reading Test (NART)* (NFER–Nelson, Windor, Connecticut, 1982).
31. Matute, H., Arcediano, F. & Miller, R. R. Test question modulates cue competition between causes and between effects *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 182–196 (1996).
32. Friston, K. J. *et al.* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1995).
33. Friston, K. J. *et al.* Spatial registration and normalisation of images. *Hum. Brain Mapp.* **2**, 165–189 (1995).
34. Talairach, J. & Tournoux, P. *Co-planar Stereotaxic Atlas of the Human Brain.* (Thieme, Stuttgart, 1988).